Education, Social Sciences, and Linguistics: Conference Series
https://journal.ummat.ac.id/index.php/ecs
Vol. 2 No. 3, October 2026, pp. 258-266

Syntactic Error Analysis of Indonesian as a Foreign Language Learners in Social Media Digital Corpora

Jastin Martino[1], Stefanus Waskito Aji Priyanka[2]
[1]Primary Teacher Education, BINUS University, Indonesia
[2]Sekolah Kristen ORA et LABORA, Indonesia
[1]Jastin.martino@binus.ac.id, [2]stefanus.priyanka@oel.sch.id

Keywords:
Indonesian as a Foreign Language (BIPA), Error Analysis, Syntactic Errors, First Language Interference, Learner Corpus, Second Language Acquisition.

ABSTRACT

Abstract: *Acquiring syntactic competence is fundamental for Indonesian as a Foreign Language (BIPA) learners, yet structural deviations remain a persistent challenge. While traditional error analysis often relies on controlled classroom data, social media provides a valuable repository of unelicited learner language to observe spontaneous interlanguage development. This study aims to analyze syntactic errors produced by BIPA learners in social media digital corpora and identify patterns of first language interference using the Error Analysis framework. Employing a descriptive qualitative design supported by quantitative statistics, an authentic corpus was purposively sampled from public social media platforms. Data processing involved AI-assisted initial categorization followed by manual researcher verification against standard Indonesian grammar. The results indicate that structural deviations are predominantly rooted in morphological (40%) and mixed (30%) error categories. The most frequent patterns involve phrase sequence inversions and morphosyntactic inaccuracies, particularly the omission or misapplication of mandatory verbal affixes. These anomalies appear driven by negative cross-linguistic transfer from learners' native typologies, alongside intralingual overgeneralization. In conclusion, the study suggests that syntactic competence in Indonesian is closely tied to morphological development. The findings support utilizing authentic digital corpora for language research and highlight the need for data-driven, form-focused BIPA instruction to address specific structural challenges.*

Kata Kunci:

Bahasa Indonesia bagi Penutur Asing (BIPA), Analisis Kesalahan, Kesalahan Sintaksis, Interferensi Bahasa Pertama, Korpus Pemelajar, Pemerolehan Bahasa Kedua.

Abstrak: Penguasaan kompetensi sintaksis merupakan hal yang fundamental bagi pemelajar Bahasa Indonesia bagi Penutur Asing (BIPA), namun penyimpangan struktur tetap menjadi tantangan yang terus dihadapi. Meskipun analisis kesalahan tradisional sering kali bergantung pada data ruang kelas yang terkontrol, media sosial menyediakan repositori bahasa pemelajar tanpa rekayasa yang berharga untuk mengamati perkembangan bahasa antara secara spontan. Penelitian ini bertujuan untuk menganalisis kesalahan sintaksis yang dihasilkan oleh pemelajar BIPA dalam digital media sosial serta mengidentifikasi pola interferensi bahasa pertama menggunakan kerangka kerja Analisis Kesalahan. Dengan menerapkan desain kualitatif deskriptif yang didukung oleh statistik kuantitatif, korpus otentik diambil secara purposive sampling dari platform media sosial publik. Pemrosesan data melibatkan pengkategorian awal berbasis AI yang dilanjutkan dengan verifikasi manual oleh peneliti terhadap tata bahasa baku bahasa Indonesia. Hasil penelitian menunjukkan bahwa penyimpangan struktur sebagian besar berakar pada kategori kesalahan morfologi (40%) dan campuran (30%). Pola yang paling sering muncul melibatkan pembalikan urutan frasa dan ketidaktepatan morfosintaksis, khususnya penghilangan atau kesalahan penerapan afiks verba yang wajib. Anomali ini tampaknya didorong oleh transfer lintas bahasa yang negatif dari tipologi bahasa ibu pemelajar, serta overgeneralisasi intralingual. Kesimpulannya, penelitian ini menunjukkan bahwa kompetensi sintaksis dalam bahasa Indonesia berkaitan erat dengan perkembangan morfologi. Temuan ini mendukung pemanfaatan korpus digital otentik untuk penelitian bahasa dan menekankan perlunya pengajaran BIPA berfokus bentuk yang berbasis data untuk mengatasi tantangan struktural secara spesifik.

Article History:

Received : 01-09-2026
Accepted : 31-10-2026

A. INTRODUCTION

The global expansion of cultural exchange, academic mobility, and bilateral economic cooperation has positioned the Indonesian language as an increasingly strategic medium for international communication (Damayanti et al., 2025). As the demographic of learners enrolled in Indonesian for Foreign Speakers (BIPA) programs continues to grow globally, a primary pedagogical imperative is equipping these individuals with functional communicative competence (Damayanti et al., 2025). Achieving this proficiency inherently requires a solid foundation in grammatical competence, with syntax functioning as the architectural core of meaning-making (Ardiansyah et al., 2026). Syntactic competence dictates the logical organization of phrases, clauses, and sentences; without it, learners struggle to map their communicative intent accurately, resulting in structural ambiguities that hinder effective discourse (Ardiansyah et al., 2026; Septiana et al., 2021). Because the Indonesian language requires strict adherence to specific structural norms, mastering its syntax is fundamental for foreign learners attempting to engage in both academic and social domains (Septiana et al., 2021).

Within the theoretical paradigms of Second Language Acquisition (SLA), the trajectory toward syntactic mastery is universally recognized as a transitional and error-prone process (Hinkel, n.d.). Moving away from behaviorist frameworks that categorized errors merely as habits requiring eradication, contemporary applied linguistics approaches learner language as a distinct, evolving cognitive system known as interlanguage (Selinker, 1972). During the acquisition process, learners continuously formulate, test, and revise hypotheses regarding the grammatical rules of the target language, making systematic errors a valuable indicator of their underlying psycholinguistic development (Corder, 1967; Hinkel, n.d.). The Error Analysis (EA) framework emerged precisely to systematically collect, document, and classify these deviations, operating under the assumption that analyzing such errors sheds light on universal areas of language difficulty (Hinkel, n.d.; Widyastuti & Noersadono, 2022).

Despite formal instruction, syntactic deviations remain a persistent challenge for BIPA learners, largely due to structural discrepancies between Indonesian and their native languages (Ardiansyah et al., 2026). Based on the Contrastive Analysis hypothesis, learners frequently experience negative cross-linguistic transfer, unconsciously mapping the syntactic rules and morphological constraints of their first language (L1) directly onto Indonesian (Fries, 1945; Odlin, 1989; Widyastuti & Noersadono, 2022). This interference manifests in various syntactic malformations, such as the misapplication of the Head-Modifier (*Diterangkan-Menerangkan*) phrase order, incorrect subject-predicate alignments, and the omission of obligatory affixes required for verb valency (Muzaki, 2022; Septiana et al., 2021). Furthermore, as learners progress, they concurrently exhibit intralingual errors, which occur when they overgeneralize newly acquired Indonesian grammatical rules or apply them across inappropriate syntactic contexts (Burhansyah, 2019; Richards, 1971). Because these combined interlingual and intralingual deviations systematically disrupt narrative coherence and chronological logic, the mechanics of syntactic interference necessitate rigorous empirical investigation (Setyaningrum et al., 2026).

A critical evaluation of the existing literature reveals profound methodological limitations in how BIPA learner interlanguage has historically been analyzed. Previously, investigations into Indonesian language acquisition have predominantly relied on highly controlled elicitation tasks, formal written essays, or direct classroom observations (Rejeki, 2015; Setyaningrum et al., 2026). While these traditional data sources provide insight into guided language production, they represent artificial environments where learners are acutely aware of grammatical monitoring, thereby obscuring spontaneous language use. Consequently, EA faced historical marginalization due to its reliance on small, decontextualized samples and the difficulty of distinguishing genuine cognitive errors from mere performance lapses (Hinkel, n.d.). Although the integration of computerized learner corpora has recently revitalized EA by enabling the large-scale analysis of systematic errors (Hinkel, n.d.), there remains a distinct scarcity of research utilizing authentic, user-generated digital corpora in the BIPA context.

Specifically, there is a remarkably limited investigation of naturally occurring learner language on social media platforms. Social media data represents a novel and expansive source of evidence that fundamentally differs from structured, traditional text corpora (Laitinen & Rautionaho, n.d.). Platforms such as YouTube provide an unmonitored environment where foreign speakers produce spontaneous, unelicited Indonesian

discourse, stripping away the artificial constraints of the classroom and exposing the raw state of their interlanguage (Muzaki, 2022). Furthermore, utilizing massive social media datasets allows researchers to overcome the elite bias inherent in traditional sampling, capturing a broader demographic of learners and providing access to language varieties and communicative strategies that are typically underrepresented in academic corpora (Laitinen & Rautionaho, n.d.). The current lack of corpus-driven error analyses utilizing social media data restricts the field's ability to systematically quantify syntactic deviations, track real-time cross-linguistic transfer, and establish robust developmental patterns across diverse L1 backgrounds.

Addressing these empirical and methodological gaps, this research offers substantial theoretical and practical contributions across multiple linguistic disciplines. Theoretically, this study advances Applied Linguistics and SLA by testing established paradigms of interlanguage and negative syntactic transfer against authentic, typologically distinct social media data. Within Corpus Linguistics, it contributes to the methodological expansion of the field by validating the reuse of unstructured, user-generated digital data to construct specialized learner corpora, thereby diversifying analytical resources and ensuring greater empirical validity (Laitinen & Rautionaho, n.d.). Practically, the findings hold transformative implications for BIPA instruction and Indonesian language pedagogy. By systematically mapping the trajectory of syntactic interference within digital discourse, educators can transition toward data-driven, form-focused pedagogy that directly targets high-frequency structural errors (Susanto, 2007). Understanding the specific syntactic vulnerabilities of learners in spontaneous settings enables the development of responsive curricula and localized teaching materials that effectively mitigate the negative transfer effects of various native languages, ultimately fostering higher levels of communicative competence.

This study aims to analyze syntactic errors produced by learners of Indonesian as a Foreign Language in social media digital corpora and to identify patterns of first language interference through the Error Analysis framework.

B. METHOD

This study employs a descriptive qualitative research design, supplemented by descriptive quantitative statistics, to systematically investigate the structural deviations produced by learners of Indonesian as a Foreign Language (BIPA). The primary theoretical framework guiding this investigation is Error Analysis, which posits that learner errors provide critical empirical evidence of the underlying cognitive processes involved in second language acquisition (Corder, 1967). To comprehensively account for the complex factors driving these grammatical deviations, the study integrates Interlanguage Theory, viewing the learner's language as a distinct, transitional linguistic system that is continuously evolving (Selinker, 1972). Furthermore, the analytical approach draws heavily upon Contrastive Analysis and cross-linguistic transfer to evaluate how the syntactic parameters of the learners' first languages interface with and subsequently disrupt their production of the target language (Gass, Behney, & Plonsky, 2020).

The empirical data for this study constitute a specialized digital learner corpus compiled from authentic, unelicited Indonesian sentences produced by foreign BIPA learners. Data were systematically collected from publicly accessible social media platforms, specifically focusing on community-driven forums such as Reddit, particularly the r/Indonesian subreddit, alongside YouTube comment sections, X, and various personal blogs or open online discussion boards. To ensure the relevance and quality of the corpus, data collection utilized purposive sampling based on strict inclusion criteria. Specifically, the selected texts had to be definitively written by foreign learners of Indonesian, contain at least one complete or attempted Indonesian sentence, and exhibit identifiable syntactic errors. Furthermore, priority was given to data where the learner's first language was explicitly stated or easily verifiable within the context of the post. To maintain the structural integrity of the dataset, duplicate sentences were rigorously removed from the final corpus, and all personal identifiers were fully anonymized.

The analytical phase of the study adhered to a systematic Error Analysis procedure adapted from foundational methodologies (Corder, 1974), which involved collecting learner sentences, identifying syntactic errors, classifying error types, interpreting potential first language interference using Contrastive Analysis, and formulating corrected sentence structures. To manage the processing of the digital text, Large Language Models were employed exclusively as analytical assistants during the initial phases of error identification and broad linguistic categorization. However, to circumvent the methodological vulnerabilities associated with automated processing, no research findings were accepted solely based on artificial intelligence output. Instead, every preliminary classification generated by the models was subjected to rigorous manual review and

verification by the human researchers. During this manual phase, the researchers revised and refined the automated analyses, subsequently mapping the identified errors against cross-linguistic typological differences to explain instances of negative syntactic transfer. Finally, the researchers provided the corrected forms for each deviant sentence strictly in accordance with Standard Indonesian Grammar, widely recognized as Tata Bahasa Baku Bahasa Indonesia (Alwi, Dardjowidjojo, Lapoliwa, & Moeliono, 2010).

To establish methodological trustworthiness, the credibility of the research findings was continuously strengthened through rigorous verification protocols. Every instance of AI-assisted classification and analysis was manually cross-referenced against the established rules of Tata Bahasa Baku Bahasa Indonesia, the theoretical principles of Error Analysis, and the contrastive frameworks governing cross-linguistic transfer. Only those linguistic analyses that survived this stringent manual verification process were retained and included in the final study. Following the qualitative interpretation, the verified dataset was summarized using descriptive quantitative statistics. Specifically, frequency counts and percentage distributions were calculated to map the prevalence of specific error categories and syntactic subtypes. This quantitative synthesis allowed the researchers to objectively rank and identify the most dominant syntactic error patterns emerging across the digital corpus.

Given the utilization of digital text, strict ethical considerations were observed throughout the data collection and analysis phases. All data were extracted exclusively from publicly accessible online domains, intentionally excluding any private messages, restricted groups, or closed conversational networks. To protect the privacy and anonymity of the internet users, all usernames, profile links, and potentially identifying personal information were completely removed prior to the commencement of the analysis. The research was conducted with the explicit understanding that the objective of analyzing social media data is to report on broader linguistic patterns, interlanguage development, and cross-linguistic phenomena, rather than to evaluate or critique individual language learners.

C. RESULT AND DISCUSSION

1. Results

The digital corpus utilized for this analysis comprises a carefully verified dataset of authentic Indonesian sentences produced by foreign learners of Indonesian as a Foreign Language. The data were extracted from publicly accessible social media platforms, including community-driven forums, video comment sections, and open online discussion boards. Through purposive sampling and a rigorous manual verification process against Standard Indonesian Grammar, a total of ten representative sentences exhibiting distinct linguistic deviations were isolated for detailed analysis. The corpus includes texts generated by learners with explicitly identified first language backgrounds, as well as learners whose native languages could not be verified from the source material. Personal identifiers were completely anonymized to maintain ethical compliance, resulting in the assignment of alphanumeric learner codes for reporting purposes. Only the verified dataset of ten sentences is included in the reported empirical results.

Table 1. Frequency Distribution of Linguistic Error Categories in the Digital BIPA Learner Corpus

Error Category	Frequency	Percentage
Morphology	4	40%
Mixed	3	30%
Syntax	2	20%
Lexicon	1	10%
Total	10	100%

The empirical results displayed in Table 1 demonstrate that morphological deviations constitute the most frequent error category in the dataset. This category accounts for four occurrences, representing exactly forty percent of the total analyzed errors. The second most prevalent category is the Mixed error classification, encompassing sentences exhibiting simultaneous deviations across multiple linguistic levels. The Mixed category appears three times, corresponding to thirty percent of the total errors. Syntactic errors rank third in frequency, with two recorded instances making up twenty percent of the overall distribution. Finally, the Lexicon category records the lowest frequency. Lexical errors appear only once, representing ten percent of the analyzed learner sentences. The distribution indicates a concentration of errors within the morphological

and mixed domains, which collectively account for seventy percent of the total observed deviations. Table 2 delineates the frequency and percentage of specific syntactic subtypes within the dataset, focusing exclusively on the five instances extracted from the Syntax and Mixed categories that contain explicitly identifiable syntactic subcategories.

Table 2. Frequency Distribution of Syntactic Error Subtypes Identified in the Digital BIPA Learner Corpus

Syntactic Subtype	Frequency	Percentage
Phrase Order (Urutan frasa)	4	80%
Conjunction Usage (Konjungsi)	1	20%
Total	5	100%

The data in Table 2 reveals that the structural errors are distributed between two specific syntactic subtypes, namely Phrase Order and Conjunction Usage. Phrase Order deviations emerge as the overwhelmingly dominant syntactic subtype within this subset. Out of the five analyzed syntactic errors, four instances are classified under the Phrase Order subtype, accounting for eighty percent of the syntactic deviations identified. In contrast, Conjunction Usage errors are significantly less frequent. Only one instance was recorded, representing twenty percent of the syntactic subtype distribution. The quantitative results highlight that the vast majority of the syntactic anomalies produced by the learners in this corpus involve the incorrect sequencing or ordering of phrases within the sentence structure, rather than the misuse of connective elements.

Table 3 provides a comprehensive overview of the representative learner sentences extracted from the digital corpus, detailing the original learner production, the classified error type, and the corresponding corrected form.

Table 3. Representative Examples of Syntactic Errors Produced by BIPA Learners in the Digital Corpus

Learner Code	Original Sentence	Error Type	Correct Form
LRN-001	Dia tidak nangis	Morphology	Dia tidak menangis.
LRN-002	Jimin adalah mahasiswa Darma Siswa paling terpandai di sini.	Morphology	Jimin adalah mahasiswa Darmasiswa terpandai di sini.
LRN-003	Aku punya banyak teman-teman	Morphology	Aku punya banyak teman.
LRN-004	Semua dosen-dosen di Jogja ini dia sangat ramah dan sangat kreatif untuk pembelajaran bahasa Indonesia.	Morphology	Semua dosen di Jogja ini sangat ramah dan sangat kreatif dalam pembelajaran bahasa Indonesia.
LRN-005	Di Mammalia, saya pilihan anjing	Mixed (Phrase Order)	Di area Mammalia, saya memilih anjing.
LRN-006	Baju baru menpakai oleh saya	Mixed (Phrase Order)	Baju baru saya pakai.
LRN-007	Buku saya sedang membaca	Mixed (Phrase Order)	Saya sedang membaca buku.
LRN-008	Mereka bahasa Inggris bagus sekali.	Syntax (Phrase Order)	Bahasa Inggris mereka bagus sekali.
LRN-009	Ronda bapak-bapak setiap malam guna untuk menjaga kebersihan dan keamanan Punden Mbah Bogo ini.	Syntax (Conjunction Usage)	Ronda bapak-bapak setiap malam untuk menjaga kebersihan dan keamanan Punden Mbah Bogo ini.
LRN-010	Saat Pabak dan ibu sedang berdansa, anak-anak bermain ikan dipelihara oleh Pabak.	Lexicon	Saat bapak dan ibu sedang berdansa, anak-anak bermain ikan yang dipelihara oleh bapak.

The representative sentences in Table 3 illustrate the practical manifestations of the quantitative data. The dataset employs anonymous learner codes, ranging from LRN-001 to LRN-010. The morphological errors documented in the corpus primarily involve the omission of mandatory prefixes, the overgeneralization of affixes, and redundant pluralization markers. The mixed errors demonstrate complex deviations where inappropriate phrase ordering is compounded by morphological inaccuracies, active-passive voice confusion, or the misplacement of predicates. Purely syntactic errors showcase clear phrase order inversion and conjunction redundancy. The single lexical error involves incorrect word choice combined with an orthographic deviation.

In summary, the empirical results reveal a total of ten verified linguistic errors within the social media digital corpus. Morphology represents the largest primary error category at forty percent of the total dataset. Among the structural deviations, phrase order constitutes the most frequent syntactic subtype, representing eighty percent of all syntactic and mixed structural errors. All learner sentences were documented exactly as produced and corrected strictly according to Standard Indonesian Grammar guidelines.

2. Discussion

The empirical evidence derived from the digital corpus demonstrates that the linguistic deviations produced by foreign learners are predominantly rooted in the complex intersection of morphology and syntax. The high prevalence of morphological and mixed errors, particularly those involving phrase sequence inversions and affixation omissions, underscores the substantial cognitive burden learners face when transitioning from their native typologies to the Indonesian grammatical system. Drawing upon the foundational Error Analysis framework established by Corder (1967), the deviations observed in this dataset represent systematic cognitive errors rather than random performance mistakes, thereby offering a direct reflection of the learners' transitional competence. Within the theoretical paradigm of Interlanguage Theory (Selinker, 1972), learners are actively formulating and testing hypotheses regarding Indonesian structural rules in a spontaneous environment. Because Indonesian relies heavily on a sophisticated agglutinative system where prefixes and suffixes strictly dictate verb valency, transitivity, and voice, learners frequently misinterpret these morphosyntactic boundaries. For instance, the omission of mandatory active prefixes or the inappropriate substitution of derivative nouns into active predicate slots illustrates a fundamental misapprehension of how Indonesian verbs operate syntactically. Based on the principles of Contrastive Analysis and the mechanisms of cross-linguistic transfer (Odlin, 1989), these recurring errors are strongly catalyzed by structural discrepancies between the learners' first languages and the target language. Learners from isolating or inflectional language backgrounds instinctively transfer their native syntactic properties, such as particle-based case markers or internal verb inflections, directly onto Indonesian base words, treating them as unbound morphemes. Furthermore, the persistent inversion of phrase sequences demonstrates profound negative syntactic transfer, where learners impose the Modifier-Head structures typical of their native languages onto the standard Indonesian Head-Modifier paradigm. Consequently, the distinct typological characteristics of Indonesian, which demand strict adherence to specific phrasal sequencing and obligatory affixation, serve as the primary catalyst for these recurring structural anomalies.

These findings strongly support and significantly extend previous empirical studies investigating BIPA learner interlanguage and syntactic interference. Recent research by Ardiansyah, Pranoto, and Damayanti (2026) identified that sentence structure and prefix application constitute the most dominant linguistic difficulties for intermediate foreign learners, a conclusion that perfectly aligns with the high frequency of phrase order and morphological deviations documented in this digital corpus. Similarly, the observation of persistent mother tongue interference driving phrase sequence errors corroborates the findings of Widyastuti and Noersadono (2022), who emphasized that the Head-Modifier construction in Indonesian consistently provokes negative transfer among learners accustomed to differing syntactic alignments. However, the current study extends the existing literature on second language writing by demonstrating the profound interdependence of syntax and morphology in spontaneous digital discourse. While previous analyses often treated syntactic and morphological errors as isolated phenomena (Muzaki, 2023), the unelicited data in this corpus reveals a high prevalence of mixed errors, indicating that morphological inaccuracies directly disrupt syntactic coherence. For example, the profound confusion between active and passive markers, such as attaching active prefixes to passive pronominal constructions, supports Wu's (2022) argument that learners' syntactic competence is inherently constrained by their morphological awareness. Furthermore, unlike traditional learner corpora collected through formal classroom assessments where the affective filter and grammatical monitor are highly active (Rejeki, 2015), this social media data highlights an abundance of intralingual overgeneralizations.

Learners frequently stack redundant conjunctions or superimpose superlative affixes alongside adverbs of degree. This specific pattern of redundancy and overgeneralization in an unmonitored digital environment reinforces James's (1998) developmental error frameworks, illustrating that unelicited second language writing exposes transitional hypotheses that formal assessments frequently obscure.

The theoretical implications of these findings offer substantial contributions to the intersecting fields of Error Analysis, Second Language Acquisition, Corpus Linguistics, and BIPA pedagogy. For Error Analysis, this study revitalizes the methodological framework by validating the use of authentic, unelicited social media data to capture the raw, unfiltered state of learner interlanguage. Historically, the discipline faced marginalization and theoretical criticism for relying on decontextualized, highly controlled classroom elicitation tasks that often obscured genuine cognitive processing errors and failed to distinguish between competence and performance (Hinkel, n.d.). By systematically analyzing spontaneous digital communication, this research demonstrates that Error Analysis remains a robust, highly relevant paradigm for mapping naturalistic language development and cross-linguistic transfer. Within the broader scope of Second Language Acquisition, the findings challenge compartmentalized models of language learning by providing empirical proof that syntactic and morphological acquisition are inextricably linked in typologically complex languages like Indonesian. The prevalence of mixed morphosyntactic errors gives profound empirical weight to the theoretical stance that syntactic competence cannot be achieved without concurrent mastery of morphological valency and word formation rules. In the realm of Corpus Linguistics, this investigation establishes the immense value of user-generated digital platforms as legitimate, dynamic repositories for specialized learner corpora. Utilizing social media data diversifies the analytical resources available to linguists, capturing spontaneous communicative strategies, bypassing the observer's paradox, and exposing the elite bias typical of traditional, highly curated academic datasets (Laitinen & Rautionaho, n.d.). For the theoretical advancement of Indonesian as a Foreign Language, the integration of authentic digital corpora reframes the fundamental understanding of interlanguage development. It portrays the acquisition of Indonesian not merely as a linear accumulation of vocabulary and isolated grammatical rules, but as a complex, dynamic process of morphosyntactic negotiation influenced heavily by both the communicative demands of digital environments and the deep-seated structural constraints of the learners' native languages.

On a practical level, the insights derived from this corpus-driven analysis necessitate a strategic recalibration of BIPA curriculum development, grammar instruction, and the design of digital language learning resources. Because morphological omissions and phrase order inversions overwhelmingly dominate the learners' spontaneous output, curriculum designers must prioritize form-focused, data-driven instruction that explicitly contrasts the structural parameters of Indonesian with the learners' native typologies (Susanto, 2007). Indonesian grammar instruction should transition away from rote vocabulary memorization toward the highly contextualized teaching of affixation, specifically emphasizing its direct impact on verb valency, transitivity, and sentence voice. Teaching materials must be systematically engineered to address the complexities of the Head-Modifier sequence and the rigid rules governing pronominal zero-passive constructions, utilizing authentic texts to demonstrate how these structures function in natural, everyday discourse. Furthermore, classroom feedback should be strategically targeted to address global morphosyntactic errors that impede communicative clarity and alter sentence logic, rather than merely correcting isolated mechanical mistakes. Educators must guide learners to recognize the profound semantic consequences of their morphological choices. In the context of modern pedagogy, these findings also underscore the urgent necessity of developing adaptive digital language learning resources. As suggested by Damayanti, Suhartono, and Savitri (2025), integrating interactive, technology-supported modules that provide immediate, constructive feedback on phrase ordering and affixation can effectively bridge the gap between grammatical theory and practical communicative competence. Ultimately, by systematically addressing these specific structural vulnerabilities through targeted pedagogy and innovative resources, educators can significantly enhance the functional proficiency of global learners.

Overall, the findings suggest that syntactic errors produced by BIPA learners are closely associated with first-language interference and developmental interlanguage processes. The use of authentic digital corpora provides valuable evidence for understanding learners' spontaneous language production and offers practical insights for improving BIPA instruction.

D. CONCLUSIONS AND SUGGESTIONS

This study analyzed syntactic errors produced by learners of Indonesian as a Foreign Language within social media digital corpora. The findings indicate that structural deviations appear to be driven by an interplay of first language interference and target-language overgeneralization. The analysis suggests that the most prominent error patterns involve phrase sequence inversions and morphosyntactic inaccuracies, particularly the omission or misapplication of mandatory verbal affixes. These structural anomalies appear to be influenced by first language interference, as learners tend to transfer native typological parameters such as isolating phrase structures or inflectional morphology—onto the Indonesian agglutinative system. Consequently, learners often experience difficulties navigating the Head-Modifier syntactic paradigm and the rules governing verb valency and voice. By utilizing unelicited social media data, this research supports the Error Analysis framework by observing learner interlanguage in spontaneous environments, which are often less monitored than formal classroom settings. The digital corpus also indicates a prevalence of intralingual developmental errors, such as the redundant stacking of conjunctions, reflecting learners' evolving hypotheses about the target language. Theoretically, the study suggests that syntactic competence in Indonesian is closely related to morphological development, supporting the use of user-generated digital text to observe naturalistic language acquisition. Practically, the findings offer empirical observations that can inform curriculum design, suggesting a need for form-focused instruction and contextualized grammar materials to address these specific structural challenges. Ultimately, the findings contribute to a better understanding of learner interlanguage in authentic digital communication and provide empirical evidence that may support future BIPA instruction and corpus-based research.

Based on the prevalence of phrase sequence inversions and morphosyntactic inaccuracies identified in this study, BIPA educators and curriculum developers are advised to refine grammar instruction to explicitly target these frequent syntactic errors. Instruction should transition from rote vocabulary memorization to form-focused pedagogy that directly contrasts Indonesian structural parameters, such as the Head-Modifier sequence and active-passive verb valency, with learners' native typologies. Furthermore, educators should actively utilize authentic learner corpora in classroom activities. By exposing students to real-world language samples and analyzing actual unelicited errors, instructors can guide learners to recognize the semantic consequences of their structural choices and correct developmental overgeneralizations.

For future researchers, it is highly recommended to expand upon this methodological framework by compiling larger and more varied datasets. Subsequent studies should include more diverse first language backgrounds to establish broader patterns of cross-linguistic transfer and verify the extent of mother-tongue interference across different language families. Additionally, researchers should expand data collection across additional social media platforms, which would help capture a wider spectrum of spontaneous interlanguage production and varying registers of unmonitored discourse.

To support these pedagogical and empirical shifts, the development of digital corpus-based BIPA learning resources must be prioritized. Curriculum designers should integrate corpus-driven teaching materials that highlight high-frequency morphosyntactic patterns and realistic communicative contexts. Developing interactive, technology-supported modules that provide immediate, data-driven feedback on phrase ordering and mandatory affixation will directly address the specific structural vulnerabilities identified in the digital corpus, thereby bridging the gap between formal grammatical theory and practical application.

Ultimately, continuing corpus-based BIPA research within authentic digital communication contexts is essential for the evolution of Indonesian language pedagogy. As digital platforms increasingly become primary domains for global interaction, analyzing unmonitored, spontaneous discourse ensures that instructional models remain empirically grounded, adaptive, and highly relevant to the actual communicative needs of modern language learners.

REFERENCES

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, A. M. (2010). Tata Bahasa Baku Bahasa Indonesia. Balai Pustaka.

Ardiansyah, R., Pranoto, A., & Damayanti, R. (2026). Linguistic difficulties in learning Indonesian as a foreign language: A syntactic and morphological analysis of foreign learners. Parole: Journal of Linguistics and Education, 16(1), 41–52. http://ejournal.undip.ac.id/index.php/parole

Burhansyah, B. (2019). Analysis of error sources in L2 written English by Indonesian undergraduate students. Studies in English Language and Education, 6(1). https://10.24815/siele.v6i1.6659

Corder, S. P. (1967). The significance of learner's errors. International Review of Applied Linguistics in Language

Teaching (IRAL), 5(1–4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Damayanti, A. M. D., Suhartono, & Savitri, A. D. (2025). BIPA learners' oral discourse on the @balaibahasaprovinsidiy YouTube channel: Implications for digital-based curriculum development. International Journal of Language Education and Cultural Review, 11(2), 214–226. https://doi.org/10.21009/ijlecr.v11i2.56191

Fries, C. C. (1945). Teaching and learning English as a foreign language. The University of Michigan Press.

Gass, S. M., Behney, J., & Plonsky, L. (2020). Second language acquisition: An introductory course. Routledge.

Hinkel, E. (n.d.). Error analysis. Seattle Pacific University.

James, C. (1998). Errors in language learning and use: Exploring error analysis. Longman.

Laitinen, M., & Rautionaho, P. (n.d.). Reuse of social media data in corpus linguistics. University of Eastern Finland.

Muzaki, H. (n.d.). Foreign speakers' Indonesian language errors on YouTube social media. Atlantis Press.

Odlin, T. (1989). Language transfer: Cross-linguistic influence in language learning. Cambridge University Press.

Rejeki, A. S. (2015). Analisis kalimat dalam teks narasi mahasiswa Program Alih Kredit Yunnan University of Nationality di Universitas Negeri Yogyakarta [Skripsi]. Universitas Negeri Yogyakarta.

Richards, J. C. (1971). A non-contrastive approach to error analysis. ELT Journal, 25, 204–219.

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching (IRAL), 10(3), 209–231. https://doi.org/10.1515/iral.1972.10.1-4.209

Septiana, H., Ardiyanti, W. N., & Dudek, S. (2021). Grammatical errors on BIPA students (Indonesian language for foreign speakers) in writing practice for beginner level. Advances in Social Science, Education and Humanities Research, 618, 159–162. Atlantis Press.

Setyaningrum, R. R., Faridah, F., Widyaiswara, T., & Ferdianto, A. V. (2026). Exploring common students error in writing recount text: A content analysis study. Jurnal Pendidikan Bahasa Inggris Proficiency, 8(1), 3–18.

Susanto, G., Pickus, D., Espree-Conaway, D., Suparmi, Russiandi, A., & Noviya, H. (2024). Indonesian language policy and perspectives on its implementation in promoting Indonesian as an international language. Cogent Arts & Humanities, 11(1). https://doi.org/10.1080/23311983.2024.2364511

Widyastuti, I., & Noersadono, A. V. (2022). Integrating contrastive analysis into classroom in English as foreign language context. Tamansiswa International Journal in Education and Science, 3(2).

Wu, M. (2022). Testing the bottleneck hypothesis: Chinese EFL learners' knowledge of morphology and syntax across proficiency levels. Second Language Research, 39(4), 1101–1124. https://doi.org/10.1177/02676583221128520